SPECIFIC AIMS

While computational techniques are currently widely used in pharmaceutical drug discovery, current generation technologies (such as docking) are unsuitable for true molecular design. Specifically, these techniques fail to to predict small molecule binding affinities to target and antitarget biomolecules with sufficient accuracy for the variety of applications currently of interest. Computational screening techniques can do better than random selection of compounds, but they lack the accuracy to guide molecular design or optimization. A new generation of physical techniques, alchemical free energy calculations, are poised to fill this void by providing a quantitative, predictive tool that can be used in multiple stages of the drug discovery pipeline, including lead optimization to improve affinity and selectivity or the retention of potency as other physical properties are optimized.

Recent success of alchemical methods sparked considerable enthusiasm, but the domain of applicability of these techniques is currently highly limited; for broad application and routine use, they need further evaluation, refinement, and development. There is a vast gulf between targets within the domain of applicability and those which are outside it. To rapidly advance, we need carefully selected systems of intermediate complexity to bridge this gulf. Without such a bridge, these techniques may encounter the same problems faced by docking and related techniques: routine failure without clear insights into why, and years to decades spent in making small methodological modifications without dramatic improvements in predictive power.

Here, we will generate data and conduct blind prediction challenges to expand the domain of applicability and robustness of these techniques. The data we generate will span the range between systems which are tractable with current methods to those only slightly less complex than pharmaceutical drug targets covered by the NIH-funded D3R effort, which runs blind challenges on protein-ligand datasets obtained from the pharmaceutical industry. This systematic set of blind prediction challenges will expand the domain applicability, pushing free energy techniques into standard application in drug design. At the same time, the data we collect will play a long lasting role in the community, going through a life cycle of collection, curation, blind challenges, and then public dissemination to serve as benchmark sets, standard reference data, and to drive construction of new and improved models. While our work here focuses primarily on generating new targeted data for a series of blind SAMPL ("Statistical Assessment of Modeling of Proteins and Ligands") challenges and running those challenges, we also plan for subsequent data dissemination. Here, we will:

Aim 1. Collect new physical property datasets to assess accuracy and spur improvements in force fields and modeling of protonation states and tautomers.
We will develop new solution-phase datasets for druglike small molecules. These data can test critical aspects of small molecule modeling (including accounting for interactions and treatment of protonation/tautomeric state) and improve our ability to predict physical properties relevant to drug discovery in new regions of chemical space. We will initially focus on aqueous/nonpolar distribution coefficients and pK_a measurements, advancing to solubilities and membrane permeabilities, using these data to drive improvements in the modeling of ligand interactions.

Aim 2. Measure affinities of drug-like compounds in supramolecular hosts to challenge quantitative models of binding in systems lacking major receptor sampling issues.
We will measure new host-guest binding free energies for cucurbiturils and deep-cavity cavitands, yielding further host-guest binding challenges which span between physical property prediction and protein-ligand binding. Host guest systems are some of the simplest cases of molecular recognition, and thus these binding data will drive improvements in modeling of simple binding systems with techniques of relevance to drug discovery.

Aim 3. Develop model protein-ligand systems that isolate specific modeling challenges of drug targets.
We will identify suitable biological protein-ligand model systems that isolate individual modeling challenges (selected to push the limits of physical techniques) and develop these for blind challenges based on new protein-ligand affinity measurements. While the initial year will feature fragment binding to human serum albumin, subsequent challenge systems will be selected using a novel informatics platform to focus on timely modeling issues.

Aim 4. Field community blind challenges to advance quantitative biomolecular design.
The data collected in Aims 1-3 will drive annual SAMPL blind challenges, allowing the field to test the latest methods and force fields to assess progress, compare them against one another head-to-head, and perform sensitivity analysis to learn how much different factors (protonation state, tautomer selection, solvent model, force field, sampling method, etc.) affect predictive power. Results will then feed back into improved treatment of these factors for subsequent challenges, driving regular cycles of application, learning, and advancement.

Overall, the data generated here and the cycles of tests in SAMPL challenges will guide new innovations in physical methods for predicting binding and physical properties, providing a foundation for the next several generations of computational methods for pharmaceutical drug discovery.